Neural networks on galaxy pattern recognition

Ricardo Contreras, Gabriel Salazar, M. Angélica Pinninghoff, Neil Nagar

University of Concepción, Chile

Keywords: Neural Networks, Galaxy classification, Galaxy Zoo Project.

Abstract

Machine learning techniques have proven to be useful in classification. In this paper we focus on the use of neural networks for the morphological classification of galaxies. We consider several parameters, e.g. color, brightness. We use them to classify galaxies into three general categories: spiral, elliptical and unknown. Initial results show neural networks can be used to classify astronomical objects.

1 Introduction

In order to study galaxies and their evolution in the universe, it is necessary to categorize them by some method. A classification scheme generally must satisfy two criteria to be successful: It should act as a shorthand means of identification of the object, and it should provide some insight to understanding the object.

Since Edwin Hubble presented in 1926 the morphological classification scheme for galaxies [9], that is now known as the Hubble tuning fork diagram because of the shape in which it is traditionally represented, as shown in Figure 1, different classification criteria have been discussed, based on particular characteristics.

Hubble's classification is based entirely on the visual appearance of a galaxy on a photographic plate. The Hubble tuning fork is a classification based on the visual appearance of the galaxies. Originally, when Hubble proposed this classification, he had hoped that it might yield deep insights, just as in the case for classifying stars a century ago. This classification scheme was thought to represent en evolutionary scheme, where galaxies start off as elliptical galaxies, then rotate, flatten and spread out as they age. Unfortunately galaxies turned out to be more complex than stars, and while this classification scheme is still used today, it does not provide us with deeper insights into the nature of galaxies.

Due to the above, the problem of classifying galaxies in a *reasonable* time and with a high degree of accuracy arises. Machine learning appears as an option to deal with this problem, for its ability to emulate the behavior of the brain, based on neuronal activity.

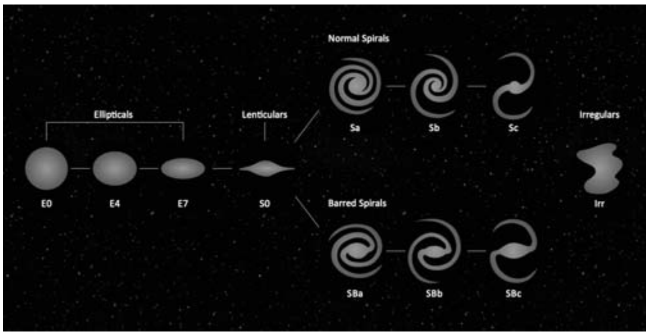

Figure 1. Hubble Sequence.

Galaxy classification allows to detect galaxy fusions and an additional set of astronomical phenomena. Galaxies classification has become a complex work because of the important volume of data generated in astronomical data centers. Some key works that consider data mining and machine learning are summarized in the following. Storrie-Lombardi [10] focused in galaxies classification depending on their morphology, according to five characteristic types (E, S0, Sa+Sb, Sc+Sd and Irr). It was used a multilayer perceptron and a set of previously classified data was used to train the network. A set of 5217 galaxies was considered, with a part (1700 galaxies) used for training. The result is 64% of properly sized galaxies.

De la Calleja and Fuentes [4] present a survey that analyzes classification processes carried out until 2004. They present an implementation which takes into account two types of learning machines. The first one deals with local regression, and the second one is a neural network with multilayer perceptron architecture. The novelty of this work is that it implements an ensemble of classifiers, i.e., it proposes a combination of classifiers as a new way of improving the performance of individual classifiers. These classifiers employ a variety of classification methodologies, and could reach different rates of correctly classified individuals. Although the authors show good results, the sample size considered is small (about 310 images).

The objective of our work is to classify galaxies, according to their morphology using neural networks. To achieve this goal, it is necessary to establish a neural net configuration that offers an improvement when compared with other similar initiatives, and to carry out a set of tests to ensure the objective is accomplished. Our work increases the number of filters considered in [11], where the galaxies were classified into three morphological types: early types (elliptic and lenticular galaxies), spirals and point sources/artifacts.

This article is structured as follows; first section consists of the current introduction. Second section is devoted to introduce the key concepts regarding both, galaxies and neural networks. Third section is devoted to the implementation issues, while section four describes the results. Finally, section 5 presents the conclusions.

2 Theoretical frame

Galaxies are sets of stars, planets, clouds of gas and dust, that inhabit the space. In the universe there exist billions of galaxies presenting different shapes and, depending on the way in which they can be visualized, astronomers classify them. Astronomer Edwin Hubble classified galaxies into four major types: spiral, barred spiral, elliptical and irregular. Most of the nearby, bright galaxies are spirals, barred spirals or ellipticals. Spiral galaxies have a bulge at the center and a flattened disk containing spiral arms. Spiral galaxies have a variety of shapes and are classified according to the size of the bulge and the tightness and appearance of the arms. The spiral arms, which wrap around the bulge, contain numerous young blue stars and lots of gas and dust. Stars in the bulge tend to be older and redder. Yellow stars like our Sun are found throughout the disk of a spiral galaxy. The disks of spiral galaxies rotate somewhat like a hurricane or a whirlpool. Barred spiral galaxies are spiral galaxies that have a bar-shaped collection of stars running across the center of the galaxy.

Elliptical galaxies do not have a disk or arms. Instead, they are characterized by a smooth, oval-shaped appearance. Ellipticals contain old stars, and possess little gas or dust. They are classified by the shape of the ball, which can range from round to oval (baseball-shaped to football-shaped). In contrast to the disks of spirals, the stars in ellipticals do not all revolve around the center in an organized way. The stars move on randomly oriented orbits within the galaxy, like a swarm of bees. Irregular galaxies are galaxies that are neither spiral nor elliptical. They tend to be smaller objects that are without definite shape, and tend to have very hot newer stars mixed in with lots of gas and dust [1].

The Sloan Digital Sky Survey or SDSS is a major multi-filter imaging and spectroscopic redshift survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory in New Mexico, United States. The project was named after the Alfred P. Sloan Foundation, which contributed significant funding.

Data collection began in 2000, and the final imaging data release covers over 35% of the sky, with photometric observations of around 500 million objects and spectra for more than 3 million objects.

2.1 Sloan Digital Sky Survey

Data release 8 (DR8), released in January 2011 [13], includes all photometric observations taken with the SDSS imaging camera, covering 14,555 square degrees on the sky (just over 35% of the full sky). Data release 9 (DR9), released to the public on 31 July 2012 [15], includes the first results from the Baryon Oscillation Spectroscopic Survey (BOSS) spectrograph, including over 800,000 new spectra. Over 500,000 of the new spectra are of objects in the Universe 7 billion years ago (roughly half the age of the universe) [14]. Data release 10 (DR10), released to the public on 31 July 2013 [16], includes all data from previous releases, plus the first results from the APO Galactic Evolution Experiment (APOGEE) spectrograph, including over 57,000 high-resolution infrared spectra of stars in the Milky Way. DR10 also includes over 670,000 new BOSS spectra of galaxies and quasars in the distant universe. The publicly available images from the survey were made between 1998 and 2009. The main goal of this project is to build a precise map of the local Universe.

The SDSS camera exhibit the capability of observe the sky from the point of view of five photometric filters (called u, g, r, i and z). It is an important issue because the classification of emitting light lines (spectrum) given by using these filters, is a standard for working in photometry (the study of the brightness in astronomical objects). The filters are based on the wavelengths that correspond to camera captured images. The curve obtained through filter u, shows the brightness intensity for the color the astronomical object exhibits in the ultraviolet portion of the emitting light. The corresponding curve for filter g shows the green color spectrum corresponding part, r filter considers the red color part of the spectrum, i filter considers the object detection in the infrared spectrum and z filter takes into account the light radiation in the near infrared spectrum. Only filters g and r show the light visible to the human eye. A detailed description on this filters system can be found in [12], [6] and [7].

2.2 Galaxy Zoo

Galaxy Zoo is the world's best-known online citizen science project, and is certainly the one with the largest number of publications based on citizen scientists input.

Volunteers vote to place astronomical objects in one of six categories: elliptical galaxies, clockwise spiral galaxies, counterclockwise spiral galaxies, edge on the spiral galaxies, stars/don't know, and mergers.

When making a classification based on a fraction of votes, there is a trade-off between the number of unclassified galaxies and the number of galaxies that are misclassified. The criterion places every galaxy in the category in which it

[1] http://hubblesite.org

received the greatest number of votes. The *clean* sample is defined by requiring 80% of the votes to be in a single category.

It all started back in July 2007, with a data set made up of a million galaxies imaged by the Sloan Digital Sky Survey, who still provide some of the images in the site today. With so many galaxies, it is assumed it would take years for visitors to the site to work through them all, but within 24 hours of launch the project received almost 70,000 classifications an hour. In the end, more than 50 million classifications were received by the project during its first year, contributed by more than 150,000 people.

That meant that many different participants saw each galaxy. This is deliberate; having multiple independent classifications of the same object is important, as it allows to assess how reliable the results are. For example, for projects that only need a few thousand galaxies but want to be sure they're all spirals before using up valuable telescope time on them, there's no problem, it is possible to use those that 100% of classifiers agree are spiral. For other projects, it may be necessary to look at the properties of hundreds of thousands of galaxies, and can use those that a majority say are spiral.

The Galaxy Zoo site relaunched with an updated design in September 2014. In this version the project combined new imaging from Sloan, giving the best ever view of the local Universe, with the most distant images yet from Hubble's CANDELS survey. The CANDELS survey makes use of the new Wide Field Camera 3, installed during the final shuttle mission to Hubble, to take ultra-deep images of the Universe. The surprise in this data was that contrary to predictions, galactic bars are still found in the most distant galaxies.

The Galaxy Zoo site relaunched in September 2014 allows more flexibility to include smaller sets of galaxies.

2.3 Data mining

In Ball el al. [1] it is presented the state of the art regarding data mining and machine learning in astronomy. This work emphasizes the advantages of using these approaches in astronomers research. Among the key issues considered it is possible to mention the following.

- Big data handling. Data mining techniques and machine learning techniques allow for working with important data volumes, a very important aspect due to the fact that day after day astronomical observatories generate huge data repositories. In the near future it will be usual talk about data volumes expressed in petabytes.

- Simplicity of algorithms. Algorithms used for this purpose have been widely used. These methods produce solution models to complex problems, non-linearities for example, often producing useful results.

- Hidden pattern detection. It is possible to obtain features that are not usually seen in a data set, as unexpected patterns. A non-supervised grouping algorithm may highlight new classes of object in a data set that may stay undetected if a previous classification is imposed by an algorithm.

- This approach can be used as a complement to traditional ways to solve these type of problems.

All these characteristics turn data mining and machine learning into interesting alternatives for helping researchers in some tasks that are very time consuming when performed manually.

2.4 Data collection

Data mining arises from the need of data collection. This task starts by recognizing existing repositories, and by foreseeing which repositories will exist in the future. Repositories in this field are called astronomical catalogues, which contain an organized set of data related to astronomical objects. These astronomical objects are usually classified in terms of common characteristics, like their morphology, origin, and the type of object among other characteristics.

Often it is necessary to pre-process data sets for the algorithmic work. The pre-processing depends on the specific problem; it is important to carefully decide the pre-processing mechanism, because it may have a serious impact in the input data. Algorithms deal with attributes, i.e., values that describe properties of an object. The most known pre-processing mechanisms are normalization and scaling.

Normalization is the process of transforming data to fit in a predetermined numeric range. This process may improve the accuracy of prediction models through the noise or nonlinearity reduction. On the other side, normalization allows for an easy identification of non-typical values, non valid values and lost values, which may produce a lost of accuracy in data.

Scaling is the process by means of which some categorical values are transformed to numerical values. For example, if it exists a category vector containing [Star, Galaxy, Quasar], it may be transformed to a numerical representation like one of the following: [1,0,0], [0,1,0] or [0,0,1] depending on the category associated to the object.

In every data set there exist some attributes that is necessary to consider, and others that do not really impact on the solution, and hence can be dismissed in future processing steps. It is clear that to consider most of attributes pertaining to a data set could lead to a poor performance for a particular algorithm. This problem is known as the dimension of punishment, due to the fact that the higher is the number of attributes in a data set, the poor is the performance because solution spaces of low density or low importance are generated. A common technique to solve this problem is known as Principal Component Analysis (PCA) [8].

2.5 Neural networks

Artificial neural networks have been used for many years as a useful tool for pattern recognition, due to the capability of processing information in a processor that contains many simple processing units called neurons.

Artificial neural networks present the following characteristics:

- They exhibit a behavior which allows to acquire knowledge through experience and training, and this knowledge is stored, into the edges that connect neurons, as a specific weight.
- They present a high change's adaptive capability, depending on the context.
- They exhibit a high fault tolerance level, in a similar way to the biological systems.
- They offer a non-linear behavior, which allows to process information supplied by other non-linear systems.

Due to the previous characterization, artificial neural networks have become a powerful mechanism for solving problems in which it is required a data processing capability that presents complex behavior [5].

On the other side, learning is the process by which a neural network modifies their weights as a response to a specific input. During the learning process, connections among neurons are modified under the action of an algorithm, which is in charge of updating the weights in different connections between neurons. A neural network is considered well trained, when weights values remain stable in time, although there are other criteria to evaluate if a neural network has successfully completed their learning stage, which depends on the specific learning algorithm.

3 Implementing the solution

For our work, it was selected the data set that corresponds to Galaxy Zoo 1, by using the release 7 of SDSS. SDSS DR7 data is stored in a database that can be accessed via the CasJobs web interface, through a query in SQL language. Parameters selected as input to the neural network include color based and brightness based parameters, profile fitting (that considers the axis ratios and log likelihoods associated with both a de Vaucouleurs and exponential fit to the two-dimensional galaxy image) and those based on the shape and the texture of a galaxy.

The design consisted in the selection of an adequate alternative for defining the neural network. The result is a neural network architecture based on the proposal of Banerji et al [11]. The selected architecture was fixed in *n : 24 : 24 : 3*, where n is the number of input parameters for every test, having two hidden layers, each one of them having 24 neurons, and an output layer with three neurons.

The neural network corresponds to a multilayer perceptron schema. The learning algorithm is Quasi-Newton, instead of Levenberg-Marquard [2], because of the high storage requirements as the number of learning patterns increases. The activating function is sigmoidal for hidden layers and the identity function for the output layer. The training process of the neural network stops when one of the following criteria is met:

Minimum gradient: 10^{-6}
Maximum of comparisons with a non-decreasing error: 10
Objective Mean Squared Error (MSE): 0
Maximum number of epochs: 1000

Input parameters were pre-processed, and every network was implemented by using MATLAB. It was necessary to dismiss galaxies containing null or spurious data. To solve the problem of filtering false positive classification it was necessary to select objects containing a voting value higher than 80% in some of the considered categories. A scaling process was applied to the vector representing an object. To make it clear, we illustrate it with an example: if a particular object is described through the vector [0.89, 0.07, 0.02, 0.02], then after the scaling process the vector changes to [1, 0, 0, 0], which is classified, according votes from Galaxy Zoo 1, as a spiral galaxy.

A test plan was established, and a set of tests was performed according to the proposed schedule. A first set of data, CLEAN, is proposed by Lintott et al. [3] in their work for presenting the Galaxy Zoo 1 catalog (it is similar to the set called Gold Sample of Banerji [11]). This set of data contains 287897 objects, being 34.2% considered spiral galaxies, 64.53% elliptical galaxies and 1.25% unknown objects or stars.

A second set of data differs from the previous set in which this set takes into account the brightest objects. This set contains 157918 astronomical objects; being 42.1% considered spiral galaxies, 56.65% elliptical galaxies and 1.25% unknown objects or stars.

4 Results

For testing, eight different test families were considered. First family: parameters based on color and profile fitting, using *u* filter. Second family: parameter based on shape and texture, using *u* filter. Third family: parameters based on color, profile fitting, shape and texture, using *u* filter. Fourth family: parameters based on color and profile fitting, using *r* filter. Fifth family: parameters based on shape, and texture, using *r* filter. Sixth family: parameters based on color, profile fitting, shape and texture, by using *r* filter. Seventh family: parameters based on color, profile fitting, shape and texture, by using *r* filter, but taking into account only the brightest objects (a set containing 157919 objects). Eighth family: parameters based on color, profile fitting and shape, using *r* filter (the difference with family seven is that the texture is not considered).

Results are summarized in Tables 1, 2 and 3, as follows. Table 1 considers families 1, 2 and 3 (type A tests); which use the *u* photometric filter information. Tests differ in the input parameters values.

u Filter	Test 1	Test 2	Test 3
Succes	94.3%	89.8%	95.9%
Error	5.7%	10.2%	4.1%
MSE	0.029	0.053	0.022

Table 1. Summary for A type

Table 2 considers families 4, 5 and 6 (type B tests); which use the r photometric filter information.

r Filter	Test 4	Test 5	Test 6
Succes	97.4%	95.6%	98.0%
Error	2.6%	4.4%	2.0%
MSE	0.013	0.023	0.011

Table 2. Summary for B type

Table 3 considers families 7 and 8 (type C tests); which are variations on the test 6 that belongs to family B.

r Filter	Test 7	Test 8
Succes	97.5%	98.0%
Error	2.5%	2.0%
MSE	0.013	0.011

Table 3. Summary for C type

From Tables shown above, it is easy to note that the results obtained by using the *r* filter show better results than those results obtained using the *u* filter. For individual tests within a group, the third test of group A presents good results, but the second test has poor performance. This indicates that, when the shape and texture parameters of a galaxy are separately considered, they do not obtain as good results as those obtained using color and profile fitting based parameters parameters for a particular astronomical object. Something similar happens for group the B tests.

From the results observed in test 7, it is observed that there is a small difference with respect to test 6. This difference can be explained by recalling that the training data set is smaller than that considered in test 6. This data set was reduced to consider only the brightest objects.

An interesting result is the one obtained in test 8. Because when the texture parameter is not considered, results do not change. In other words, according to the results, when the *r* filter is used, the texture does not affect the classification process.

5 Conclusions

In this work, we used neural networks to perform morphological classification of galaxies having as reference the Galaxy Zoo catalog. The results show that using the appropriate photometric filter, and carefully choosing the input parameters, the resulting classification offers interesting results (on average 90% of success).

In general, the results associated with the *r* filter are better than the results obtained with the *u* filter. This was to be expected, because the *r* filter shows light which is visible to the human eye. This is the same distinguishing feature used in the Galaxy Zoo catalog.

An interesting fact is that the neural networks which use parameters based on color and the profile fitting, behave better that neural networks that use parameters based on shape and texture.

Acknowledgements

This research was partially supported by FONDEF Conicyt, Project D11/1060.

References

[1] N. Ball and R. Brunner. Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics*, 2010.

[2] C. Bishop. Neural Networks for Pattern Recognition. , 1995.

[3] K. Schawinnski et al. C. Lintott. Galaxy Zoo 1: Data Release of Morphological Classifications for nearly 900,000 galaxies. *Monthly Notices of the Royal Astronomic Society*, 2010.

[4] J. de la Calleja and J. Fuentes. Machine learning and image analysis for morphological galaxy classification. *Monthly Notices of the Royal Astronomical Society*, 2004.

[5] S. Haykin. Neural Networks. , 1999.

[6] D. Tucker et al. J. Smith. The u' g' r' i' z' Standard-Star System. *The Astronomical Journal*, 2002.

[7] D. Tucker et al. J. Smith. Historical View of the u' g' r' i' z' Standard System. *ASP Conference Series*, 2007.

[8] I.T. Jollife. Principal Component Analysis. , 2002.

[9] K. Masters et al. L. Fortson. Galaxy Zoo: Morphological Classification and Citizen Science. *Advances in Machine Learning and Data Mining for Astronomy*, 2011.

[10] O. Lahav et al. LM.C. Storrie-Lombardi. Morphological Classification of Galaxies by Artificial Neural Networks. *Monthly Notices of the Royal Astronomical Society*, 1992.

[11] O. Lahav et al. M. Banerji. Galaxy Zoo: Reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society*, 2010.

[12] T. Ichikawa et al. M. Fukugita. The Sloan Digital Survey Photometric System. *The Astronomical Journal*, 1996.

[13] SDSS2011. SDSS Data Release 8. sdss3.org. Retrieved 2011-01-10. , 2011.

[14] SDSS2012. http://www.nyu.edu/about/news-publications /news/2012/08/08/new-3d-map-of-massive-galaxies-and-black-holes-offers-clues-to-dark-matter-dark-energy.html. , 2012.

[15] SDSS2012. SDSS Data Release 8. sdss3.org. Retrieved 2012-07-31. , 2012.

[16] SDSS2013. SDSS Data Release 8. sdss3.org. Retrieved 2013-08-04. , 2013.